# Postprint of Applied Research on Optimizing Statistical Modeling Strategies for Genetic Risk of Mild Cognitive Impairment Based on metaPRS and APOE 4

**Authors:** Zimeng Li, Rong Wang, Shuai Chen, Caili Zhao, Xiaocong Wang, Yalu Wen, Long Liu

**Date:** 2022-10-31T00:00:00+00:00

## Abstract

Background Mild cognitive impairment (MCI) represents a critical window for intervention and delaying the progression of dementia. Prior research has demonstrated that MCI exhibits a strong association with genetic factors, and apolipoprotein E (APOE) 4 is a widely recognized risk allele for MCI in the medical community. Owing to the absence of summary-level data from genome-wide association studies (GWAS) on MCI, the current common practice is to utilize GWAS summary data from Alzheimer' s disease (AD) as the base dataset to calculate polygenic risk scores (PRS) for MCI, leading to unsatisfactory performance in genetic risk prediction.

Objective This study utilizes meta-polygenic risk score (metaPRS) and APOE 4 as important predictive factors to investigate and optimize statistical modeling strategies for genetic risk of MCI from the perspectives of generalized linear models and machine learning.

Methods Twelve sub-phenotype PRS for MCI were calculated and integrated into metaPRS for MCI using an elastic net Logistic regression model. The age-adjusted APOE 4 weighted sum (SCOREAPOE) was computed using age-corrected APOE 4 effect sizes. Various strategies for incorporating predictive factors were developed based on metaPRS, SCOREAPOE, and basic demographic information (age, sex, education level), using XGBoost, GBM, Logistic regression, and Lasso regression as statistical modeling approaches. Prediction performance of genetic risk statistical modeling for MCI was assessed using AUC and F-measure.

Results metaPRS and SCOREAPOE exhibit high predictive value for genetic risk of MCI. Following the incorporation of metaPRS, SCOREAPOE, and

basic demographic information (age, sex, education level), the prediction performance of each statistical modeling method was as follows: XGBoost (AUC=0.69, F-measure=0.88), GBM (AUC=0.76, F-measure=0.87), Logistic regression (AUC=0.77, F-measure=0.89), Lasso regression (AUC=0.76, F

## Full Text

## Preamble

### Application of metaPRS and APOE 4 to Optimize Genetic Risk Statistical Modeling Strategies for Mild Cognitive Impairment

Li Zimeng[1], Wang Rong[1], Chen Shuai[1], Zhao Caili[1], Wang Xiaocong[3], Wen Yalu[1],[2], *Liu Long*[1],[2]

[1]Department of Health Statistics, School of Public Health, Shanxi Medical University, Taiyuan 030000, China
[2]Shanxi Key Laboratory for Risk Assessment of Major Diseases, Taiyuan 030000, China
[3]School of Public Health and Preventive Medicine, Monash University, VIC 3800, Australia

*Corresponding authors:
Wen Yalu, Professor, Doctoral Supervisor; Email: wenyalu1031shanxi@126.com
Liu Long, Lecturer, Master Supervisor; Email: biostat-ll@sxmu.edu.cn

#### Abstract

**Background** Mild cognitive impairment (MCI) represents a critical window for intervention to delay dementia progression. Previous studies have demonstrated a strong association between MCI and genetic factors, with Apolipoprotein E (APOE) 4 recognized as a significant risk allele. However, due to the absence of publicly available genome-wide association study (GWAS) summary statistics for MCI, researchers commonly substitute Alzheimer' s disease (AD) GWAS data as the base dataset for calculating polygenic risk scores (PRS), resulting in suboptimal predictive performance for MCI genetic risk.

**Objective** This study employs meta-polygenic risk score (metaPRS) and APOE 4 as key predictors to explore and optimize statistical modeling strategies for MCI genetic risk from both generalized linear model and machine learning perspectives.

**Methods** We calculated 12 sub-phenotype PRSs for MCI and integrated them into a metaPRS using an elastic-net logistic regression model. An age-adjusted APOE 4 effect size was used to compute the weighted APOE 4 sum score (SCOREAPOE). Different predictor inclusion strategies based on metaPRS, SCOREAPOE, and basic demographic information (age, sex, education level) were evaluated using XGBoost, GBM, logistic regression, and Lasso regression. Model performance was assessed using AUC and F-measure.

**Results** Both metaPRS and SCOREAPOE demonstrated high predictive value for MCI genetic risk. When metaPRS, SCOREAPOE, and demographic variables were included, the predictive performance was: XGBoost (AUC=0.69, F-measure=0.88), GBM (AUC=0.76, F-measure=0.87), logistic regression (AUC=0.77, F-measure=0.89), and Lasso regression (AUC=0.76, F-measure=0.92).

**Conclusion** With moderate sample sizes (<500), the combination of metaPRS, SCOREAPOE, and demographic predictors using Lasso regression yielded the best performance for MCI genetic risk prediction, offering novel insights for statistical modeling of genetic risk in MCI and other complex diseases.

**Keywords** Mild cognitive impairment; Polygenic risk score; MetaPRS; APOE 4; Genetic risk prediction; Statistical modeling optimization

Mild cognitive impairment (MCI) is a critical stage for intervention and delaying dementia progression [1]. Research indicates that MCI results from combined genetic and environmental factors, with Apolipoprotein E (APOE) 4 showing strong association [2]. Polygenic risk score (PRS) is a widely used method for predicting genetic risk in complex diseases. Due to the unique disease status of MCI, no publicly available international GWAS summary data exist for MCI. Consequently, AD GWAS summary data are commonly used as the base dataset for MCI PRS calculation, leading to unsatisfactory prediction performance with AUC values typically ranging from 0.58 to 0.68 [3]. Abraham et al. [4] proposed meta-polygenic risk score (metaPRS), which enhances prediction accuracy by effectively integrating multiple sub-phenotype PRSs. MetaPRS has demonstrated excellent performance in ischemic stroke, depression, and coronary artery disease. Furthermore, studies show that basic demographic information (age, sex, education level) [5] and the weighted APOE 4 sum score (SCOREAPOE) [6] have substantial predictive value for MCI, warranting further investigation.

Statistical modeling methods for MCI genetic risk primarily include generalized linear models (GLM) and machine learning (ML). Effective prediction modeling for complex diseases requires two key features: the ability to handle non-normally distributed phenotypes and to address potential complex functional relationships among predictors. Lasso regression, a GLM method using L1 regularization, offers greater sparsity than logistic regression, enabling selection of important predictors with strong model interpretability. In contrast, ML methods like XGBoost (Extreme Gradient Boosting) and GBM (Gradient Boosting Machine) combine multiple weak supervised models into robust strong models, better capturing complex variable relationships, though typically with lower interpretability than GLM.

This study employs metaPRS, SCOREAPOE, and demographic information as predictors for MCI genetic risk modeling. Considering potential complex functional relationships among these predictors and the intricate phenotypic data characteristics, we evaluated XGBoost, GBM, logistic regression, and Lasso regression to explore and optimize MCI genetic risk modeling strategies, providing

new perspectives and scientific evidence for high-risk population identification, early prevention, intervention, and precision medicine research in MCI.

### 1.1.1 Data Sources

Genomic data for this MCI genetic risk prediction study were obtained from the UK Biobank (UKB) and the Alzheimer' s Disease Neuroimaging Initiative (ADNI). UKB is a large prospective cohort study and biomedical database containing cognitive function tests, blood pressure measurements, anthropometric data, blood test results, genetic sequencing data, whole-body imaging (e.g., brain and cardiac MRI), and follow-up data. ADNI is a large-scale cohort study collecting demographic variables (age, sex, education level), brain imaging data, biomarkers, and genetic sequencing data.

This study focused on brain structural imaging phenotypes, selecting four primary brain tissue structures: white matter volume, grey matter volume, cerebrospinal fluid (CSF) volume, and total brain volume. Additionally, we included brain structural imaging phenotypes previously associated with MCI: white matter hyperintensities (WMH), pallidum, caudate, hippocampus, amygdala, accumbens, putamen, and thalamus volumes [7, 8].

### 1.1.2 Quality Control

Since the UKB database comprises exclusively white participants, we retained only non-Hispanic whites from ADNI after principal component analysis (see Appendix Materials 1-3, http://cstr.cn/31253.11.sciencedb.j00150.00009) to control for population stratification and ensure demographic similarity between databases. PLINK 1.9 was used to exclude individuals with missing rates >10% and SNPs with genotype missing rates >10%. Post-quality-control, SNPs common to both UKB and ADNI were extracted by physical position. The final dataset included 488,371 individuals with 694,020 SNPs in UKB for GWAS analysis of each sub-phenotype, and 325 individuals with 694,020 SNPs in ADNI.

## 1.2 Methods

The study design comprised three stages (Figure 1 [Figure 1: see original paper]). Stage 1: Calculated 12 sub-phenotype PRSs for MCI in the ADNI dataset. Stage 2: Integrated the 12 sub-phenotype PRSs using elastic-net logistic regression to compute metaPRS. Stage 3: Validated different predictor inclusion strategies and modeling methods via 10-fold cross-validation.

### 1.2.1 Genome-Wide Association Study

GWAS performs population-level statistical analysis of single nucleotide polymorphisms (SNPs) to identify and characterize associations between SNPs and disease progression or outcomes [9]. Results are visualized using Quantile-Quantile (Q-Q) plots and Manhattan plots, with Manhattan plots displaying

SNP significance levels and Q-Q plots showing the relationship between expected and observed test statistics. The lambda statistic assesses whether principal components are needed to control population stratification [9].

### 1.2.2 metaPRS Construction

(1) Sub-phenotype PRSs were calculated using the classic clumping and thresholding (C+T) method. The PRS formula is: = βj Xij, where represents the -th individual, indexes each SNP, $\beta$ represents the effect size from GWAS summary data, and Xij is the number of risk alleles for SNP in individual .

(2) In the ADNI database (n=325), 30% of individuals were randomly selected. An elastic-net logistic regression model integrated the 12 sub-phenotype PRSs, with coefficients ($ $1, , $\beta$ ) from the final model serving as weights [4, 10] to construct the metaPRS prediction model.

(3) Sub-phenotype PRS-level weights were converted to SNP-level weights using $\beta$ _ = $\alpha$ 1 + + $\alpha$ , where 1, , are standard deviations of each sub-phenotype PRS in the training set, and $\alpha$ 1, , $\alpha$ are effect sizes of the -th SNP' s allele for each sub-phenotype. If a SNP was not included in the -th score, its effect size $\alpha$ was set to 0.

(4) metaPRS was calculated as: metaPRS = βsnp_i × Ni, where βsnp_i is the effect size of the -th SNP and Ni is the number of effect alleles carried for SNP .

### 1.2.3 Predictor Inclusion Strategies

Predictor inclusion strategies were constructed based on demographic and genetic information. Since rs429358 is the most significant locus in the APOE 4 linkage disequilibrium region, it was selected to represent APOE 4 [11]. As APOE 4 allele frequency varies with age [12], we calculated age-adjusted APOE 4 effect sizes using $\beta$ 4 = ln (age $ $60: $\beta$ 4 = 0.542; 60<age$ $70: $\beta$ 4 = 0.419; 70<age$ $80: $\beta$ 4 = 0.577; age>80: $\beta$ 4 = 0.425 [13]) and computed the weighted APOE 4 sum score [6] as: − = $\beta$ , where indexes the -th individual, $\beta$ is the APOE 4 effect size, and is the number of risk alleles at rs429358. The predictor inclusion strategies are summarized in Table 1 .

**Table 1** Description of predictor inclusion strategies for statistical modeling of genetic risk for MCI

| Predictor Inclusion Strategy | Variables |
|---|---|
| Age + Sex + Education + PRSpheno_{12} | Demographics + 12 sub-phenotype PRSs |

| Predictor Inclusion Strategy | Variables |
|---|---|
| Age + Sex + Education + PRSpheno_{12} + SCOREAPOE | Demographics + 12 sub-phenotype PRSs + APOE 4 weighted sum |
| Age + Sex + Education + metaPRS | Demographics + metaPRS |
| Age + Sex + Education + metaPRS + SCOREAPOE | Demographics + metaPRS + APOE 4 weighted sum |

*Note: PRSpheno_{12} = 12 sub-phenotype PRSs constructed from UKB GWAS summary data; metaPRS = metaPRS integrated from 12 MCI sub-phenotype PRSs; SCOREAPOE = weighted sum score for APOE 4.*

### 1.2.4 Statistical Modeling Methods

(1) **XGBoost** (Extreme Gradient Boosting) is an ensemble learning algorithm [14] that utilizes second-order derivative information to train tree models, with tree complexity incorporated as a regularization term to enhance generalization. The objective function is: ( ) = ( ˆ, ) + Ω( ). The loss function is ( ˆ, ), and the regularization term is Ω( ) = γT + ‖ ‖2, where T represents the number of leaf nodes and denotes leaf node scores. Smaller regularization values indicate lower complexity and stronger generalization.

(2) **GBM** (Gradient Boosting Machine) is a commonly used ML algorithm comprising numerous simple decision trees that iteratively learn residuals to reduce loss function values, offering high interpretability [15]. GBM can model relationships between phenotypes and predictors without prior data structure assumptions, demonstrating strong generalization ability. GBM is expressed as an additive regression model: ∗ = 1 + ( ∗; ) + , where ∗ is the phenotype, X represents predictors, is the residual, and controls variance subtraction from residuals at each iteration, balancing model number and predictor correlation. Smaller values require more combined models to achieve the same training error rate but yield better validation performance.

(3) **Logistic Regression** is the most common statistical model for binary outcomes, with the general form: Logit(P) = Log ( 1 − P ) = a + b1x1 + + bmxm, where 1, , are predictors and 1, are regression coefficients. Through simple transformation, the predicted event probability is: P = exp(a+b1x1+ +bmxm) 1+exp(a+b1x1+ +bmxm).

(4) **Lasso Regression**, proposed by Tibshirani in 1997 [16], constructs optimal penalized linear models. Strong penalization drives some predictor coefficients toward zero, eliminating predictors with zero coefficients from the model. Lasso regression effectively produces sparse coefficient vectors, selects informative features, and achieves superior model performance.

### 1.3 Statistical Analysis

All analyses were performed using R software (version 4.1.0). The XGBoost, gbm, stats, and glmnet packages were used for XGBoost, GBM, logistic regression, and Lasso regression, respectively. All prediction models were validated using 10-fold cross-validation, with performance evaluated by F1 score (F-measure) and AUC. F-measure is a reliability metric for binary classification models; higher values indicate better balance between precision and recall and higher model reliability.

## Results

### 2.1 Study Subject Characteristics

The MCI group had a mean age of (70.66$±7.00)$years, $whilethecontrolgroupmeanagewas$(74.26±$5.69) years. APOE 4 allele frequency was 45.79% in the MCI group and 27.93% in the control group (Table 2 ).

**Table 2** General characteristics of 325 ADNI participants

| Characteristic | Cognitively Normal (N=111) | MCI (N=214) |
|---|---|---|
| Age (years) | 74.26$±5.69\|70.66±7.00\|\|$Sex$(Male/Female)\|59/52\|114/100\|\|$Education$(years)\|16.42$± | |
| APOE 4 allele | 31 (27.93%) | 98 (45.79%) |

### 2.2 Genome-Wide Association Study

Lambda statistics for all 12 sub-phenotypes were close to 1, indicating appropriate adjustment for population stratification (Figure 2 [Figure 2: see original paper]). Manhattan plots revealed SNPs reaching Bonferroni significance (p<5$×10$^{\{-8\}}$, firsthorizontalline)foramygdala, caudate, CSF, pallidum, putamen, andWMHphenotypes, withtheseSNPslocatedin 6}$ (second horizontal line).

The p<5$×10^{-8}$ threshold is reliable, as no SNPs at this level have been proven false positive [18]. Reed [9] identified significant SNPs at p<5$×10$^{\{-6\}}$, alessstringentthresholdrequiringfurthervalidation, similartoEdmondson′sapproach[19].Therefore, weadoptedbothB 8})$andBonferronithreshold$(p < 5×10^{-6}$) to identify informative SNPs across sub-phenotype GWAS summary data.

### 2.3 metaPRS Construction

Pearson correlation coefficients between predictors are shown in Figure 3 [Figure 3: see original paper]. Notable correlations included: PRSHippocampus and metaPRS (r=-0.6), PRSWMH and metaPRS (r=0.5), PRSPallidum and metaPRS (r=-0.5), PRSCSF and PRSAccumbens (r=-0.4), PRSCSF and PRSTotal brain (r=-0.4), PRSTotal brain and PRSGrey matter (r=-0.4), and PRSAccumbens and PRSThalamus (r=0.4).

### 2.4 Validation of Predictor Inclusion Strategies

Comparisons between Strategy I vs. II (Group A) and Strategy III vs. IV (Group B) showed that strategies incorporating SCOREAPOE consistently outperformed those without SCOREAPOE, confirming the predictive value of APOE 4 for MCI. Group C (Strategy II vs. IV) demonstrated that Strategy IV (metaPRS-based) outperformed Strategy II (12 sub-phenotype PRS-based) across all four statistical modeling methods, indicating that metaPRS-optimized predictor inclusion strategies are superior (Figure 4 [Figure 4: see original paper]).

### 2.5 Evaluation of Statistical Modeling Performance

Overall, Lasso regression demonstrated superior predictive performance compared to the other three methods. In Group A, Lasso regression achieved higher F-measure values across different predictor inclusion strategies. Under Strategy IV (metaPRS + SCOREAPOE), F-measure values were: XGBoost (0.88), GBM (0.87), logistic regression (0.89), and Lasso regression (0.92). In Group B, AUC distributions across methods under Strategy IV were similar, with median values: XGBoost (0.69), GBM (0.76), logistic regression (0.77), and Lasso regression (0.76) (Figure 5 [Figure 5: see original paper]).

## Discussion

This study explored and constructed optimal statistical modeling strategies for MCI genetic risk prediction using 12 sub-phenotype PRSs, metaPRS, SCOREAPOE, and demographic information as predictors, with XGBoost, GBM, logistic regression, and Lasso regression as modeling methods. Our findings demonstrate that metaPRS and SCOREAPOE have high predictive value for MCI genetic risk, and Lasso regression is an ideal method for MCI genetic risk modeling when sample sizes are modest (<500).

Age-adjusted APOE 4 effect sizes significantly improved MCI prediction when incorporated as weighted scores, underscoring the importance of SCOREAPOE. Previous research indicates that APOE 4 allele frequency declines with age and that its effect size is age-dependent [12]; our study validates the rationale and scientific merit of using age-adjusted APOE 4 effect sizes as independent predictors. Additionally, metaPRS-based predictor inclusion strategies outperformed

both 12 sub-phenotype PRS-based strategies and previous MCI prediction approaches. The combination of metaPRS and SCOREAPOE surpassed three alternative predictor strategies (Figure 4). Prior MCI predictions using AD GWAS data achieved AUCs of 0.58-0.68 [3] because those GWAS summary data represented AD binary outcomes. In contrast, our study selected 12 MCI-related brain imaging phenotypes, reasonably integrated correlated sub-phenotype PRSs into metaPRS, and compared XGBoost, GBM, logistic regression, and Lasso regression, ultimately achieving higher model performance. Future MCI genetic risk research should focus on identifying relevant predictors and developing integration methods. While our prediction model has not yet reached clinical diagnostic standards, it represents significant progress over previous studies.

Considering both F-measure and AUC, Lasso regression showed the best performance. First, in Strategies I (12 sub-phenotype PRSs) and II (12 sub-phenotype PRSs + SCOREAPOE), Lasso regression outperformed other methods, primarily due to its stronger ability to produce sparse coefficient vectors, making penalized linear regression more suitable for genetic risk prediction models built from correlated sub-phenotypes. Second, in Strategies III (metaPRS) and IV (metaPRS + SCOREAPOE), XGBoost underperformed relative to the other three methods, likely because our sample size was insufficient for XGBoost to demonstrate its advantages over Lasso regression. Christodoulou et al. [20] reviewed 75 studies (median sample size=1,250, range=72-3,994,872) and found no performance advantage of ML over logistic regression for clinical prediction models. Other studies [21] have shown that while XGBoost performs best among ML methods (naive Bayes, XGBoost, SVM, etc.), its performance heavily depends on sample size, offering no clear advantage when $n<500$.

The relatively small training sample size may limit generalizability. Additionally, using SNPs at common physical positions from UKB and ADNI may have omitted MCI-relevant genetic information. Future studies should consider measuring rare variants. Moreover, with only four statistical modeling methods examined, further exploration of alternative approaches to improve MCI genetic risk prediction accuracy and development of novel statistical models are warranted.

In summary, the statistical modeling strategy using metaPRS, SCOREAPOE, and demographic information (age, sex, education) as predictors with Lasso regression achieved favorable predictive performance, providing scientific evidence for precision medicine and early intervention in MCI. Integrating MCI genetic risk prediction into routine health screenings could substantially improve detection rates, enabling early intervention and reducing disease burden on families and society.

**Author Contributions**: Li Zimeng conceptualized the study, performed feasibility analysis, interpreted results, and drafted/revised the manuscript. Wang Rong, Chen Shuai, and Zhao Caili collected, translated, and organized literature/materials. Wang Xiaocong collected data. Wen Yalu and Liu Long pro-

vided core supervision and overall responsibility for the article. All authors approved the final manuscript.

**Conflict of Interest**: The authors declare no conflicts of interest.

**References**

[1] ANDERSON N D. State of the science on mild cognitive impairment (MCI) [J]. CNS spectrums, 2019, 24(1): 78-87.

[2] LUO Y, TAN L, THERRIAULT J, et al. The Role of Apolipoprotein E 4 in Early and Late Mild Cognitive Impairment [J]. European Neurology, 2021, 84(6): 472-480.

[3] LEONENKO G, SHOAI M, BELLOU E, et al. Genetic risk for alzheimer disease is distinct from genetic risk for amyloid deposition [J]. Annals of neurology, 2019, 86(3): 427-435.

[4] ABRAHAM G, MALIK R, YONOVA-DOING E, et al. Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke [J]. Nature communications, 2019, 10(1): 1-10.

[5] RITCHIE K. Mild cognitive impairment: an epidemiological perspective [J]. Dialogues in clinical neuroscience, 2004, 6(4): 401-408.

[6] LEONENKO G, BAKER E, STEVENSON-HOARE J, et al. Identifying individuals with high risk of Alzheimer' s disease using polygenic risk scores [J]. Nat Commun, 2021, 12(1): 4506.

[7] VAN DEN BERG E, GEERLINGS M I, BIESSELS G J, et al. White matter hyperintensities and cognition in mild cognitive impairment and Alzheimer' s disease: a domain-specific meta-analysis [J]. Journal of Alzheimer' s disease, 2018, 63(2): 515-527.

[8] ZACKOVá L, JáNI M, BRáZDIL M, et al. Cognitive impairment and depression: Meta-analysis of structural magnetic resonance imaging studies [J]. NeuroImage: Clinical, 2021, 32: 102830.

[9] REED E, NUNEZ S, KULP D, et al. A guide to genome-wide association analysis and post-analytic interrogation [J]. Statistics in medicine, 2015, 34(28): 3769-3792.

[10] 牛晓歌. 基于大型前瞻性队列构建和评价中国人群脑卒中多基因遗传风险评分 [D]; 北京协和医学院, 2021.

[11] ANDREWS S J, FULTON-HOWARD B, GOATE A. Interpretation of risk loci from genome-wide association studies of Alzheimer's disease [J]. The Lancet Neurology, 2020, 19(4): 326-335.

[12] BELLOU E, BAKER E, LEONENKO G, et al. Age-dependent effect of APOE and polygenic component on Alzheimer' s disease [J]. Neurobiology of aging, 2020, 93: 69-77.

[13] BONHAM L W, GEIER E G, FAN C C, et al. Age-dependent effects of APOE epsilon4 in preclinical Alzheimer' s disease [J]. Ann Clin Transl Neurol, 2016, 3(9): 668-677.

[14] CHEN T, GUESTRIN C. XGBoost: A Scalable Tree Boosting System. KDD' 16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2016: 785-794 [Z]. 2016.

[15] EATON J E, VESTERHUS M, MCCAULEY B M, et al. Primary sclerosing cholangitis risk estimate tool (PREsTo) predicts outcomes of the disease: a derivation and validation study using machine learning [J]. Hepatology, 2020, 71(1): 214-224.

[16] TIBSHIRANI R. The lasso method for variable selection in the Cox model [J]. Statistics in medicine, 1997, 16(4): 385-395.

[17] LI J, LU Q, WEN Y. Multi-kernel linear mixed model with adaptive lasso for prediction analysis on high-dimensional multi-omics data [J]. Bioinformatics, 2020, 36(6): 1785-1794.

[18] DUDBRIDGE F, GUSNANTO A. Estimation of significance thresholds for genomewide association scans [J]. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society, 2008, 32(3): 227-234.

[19] EDMONDSON A C, BRAUND P S, STYLIANOU I M, et al. Dense genotyping of candidate gene loci identifies variants associated with high-density lipoprotein cholesterol [J]. Circulation: Cardiovascular Genetics, 2011, 4(2): 145-154.

[20] CHRISTODOULOU E, MA J, COLLINS G S, et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models [J]. Journal of clinical epidemiology, 2019, 110: 12-22.

[21] RáCZ A, BAJUSZ D, HéBERGER K. Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification [J]. Molecules, 2021, 26(4): 1111.

**Data Availability Statement**: The scientific data supporting this study have been publicly released in the Science Data Bank of the Chinese Academy of Sciences and can be accessed at http://cstr.cn/31253.11.sciencedb.j00150.00009, DOI: 10.57760/sciencedb.j00150.00009, CSTR: 31253.11.sciencedb.j00150.00009.

*Note: Figure translations are in progress. See original paper for figures.*